Machine Learning 1

Lecture 3.2 - Supervised Learning
Linear Regression via Maximum Likelihood
Optimization

Erik Bekkers

(Bishop 3.1.1)

Slide credits: Rianne van den Berg

Image credit: Kirillm | Getty Images

Linear Regression

- Regression: $D = \{(\mathbf{x}_1, t_1), ..., (\mathbf{x}_N, t_N)\}$
  - Input variables $\underline{x}_i \in \mathbb{R}^D$
  - Target variables $t_i \in \mathbb{R}$

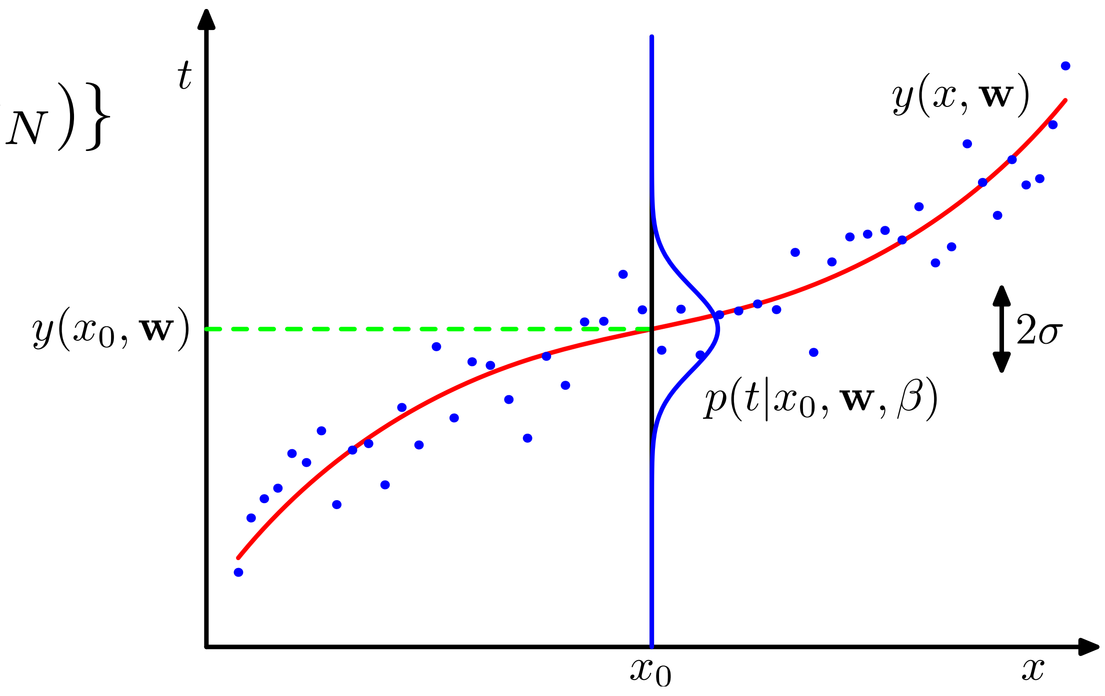


Figure: Gaussian conditional distribution (Bishop 1.16)

- Linear model with basis functions

$$y(\mathbf{x}, \mathbf{w}) = \underline{w}^T \underline{\phi}(\underline{x})$$

$$\underline{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{pmatrix} \in \mathbb{R}^M$$

$$\underline{\phi}(\underline{x}) = \begin{pmatrix} 1 \\ \phi_1(\underline{x}) \\ \phi_2(\underline{x}) \\ \vdots \\ \phi_{M-1}(\underline{x}) \end{pmatrix} \in \mathbb{R}^M$$

Maximum Likelihood

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

- Assume gaussian noise around the target

$$t = y(\underline{x}, \underline{w}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \beta^{-1})$$

- $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t \mid \underbrace{\underline{w}^T \underline{\phi}(\underline{x})}_{y(\underline{x}, \underline{w})}, \beta^{-1})$

- Dataset: $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and $\mathbf{t} = (t_1, ..., t_N)^T$

data matrix ↑ $D \times N$ vector of size N

- Likelihood function

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \sqrt{\frac{\beta}{2\pi}} \, e^{-\frac{\beta}{2}(t_i - \underline{w}^T \underline{\phi}(x_i))^2}$$

ML: Sum-of-Squares Error

- Likelihood: $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i), \beta^{-1})$

- Log likelihood $\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) =$

$$\frac{N}{2} \log \beta - \frac{N}{2} \log 2\pi - \frac{\beta}{2} \sum_{i=1}^{N} (t_i - \underline{w}^T \underline{\phi}(\underline{x}_i))^2$$

- Sum-of-squares error: $E_D(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} (t_i - \underline{w}^T \underline{\phi}(\underline{x}_i))^2$

- For comparison of different dataset sizes *N*

$$E_D^{\mathrm{RMSE}}(\mathbf{w}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (t_i - \underline{w}^T \underline{\phi}(\underline{x}_i))^2}$$

Example: Sum-of-Squares Error

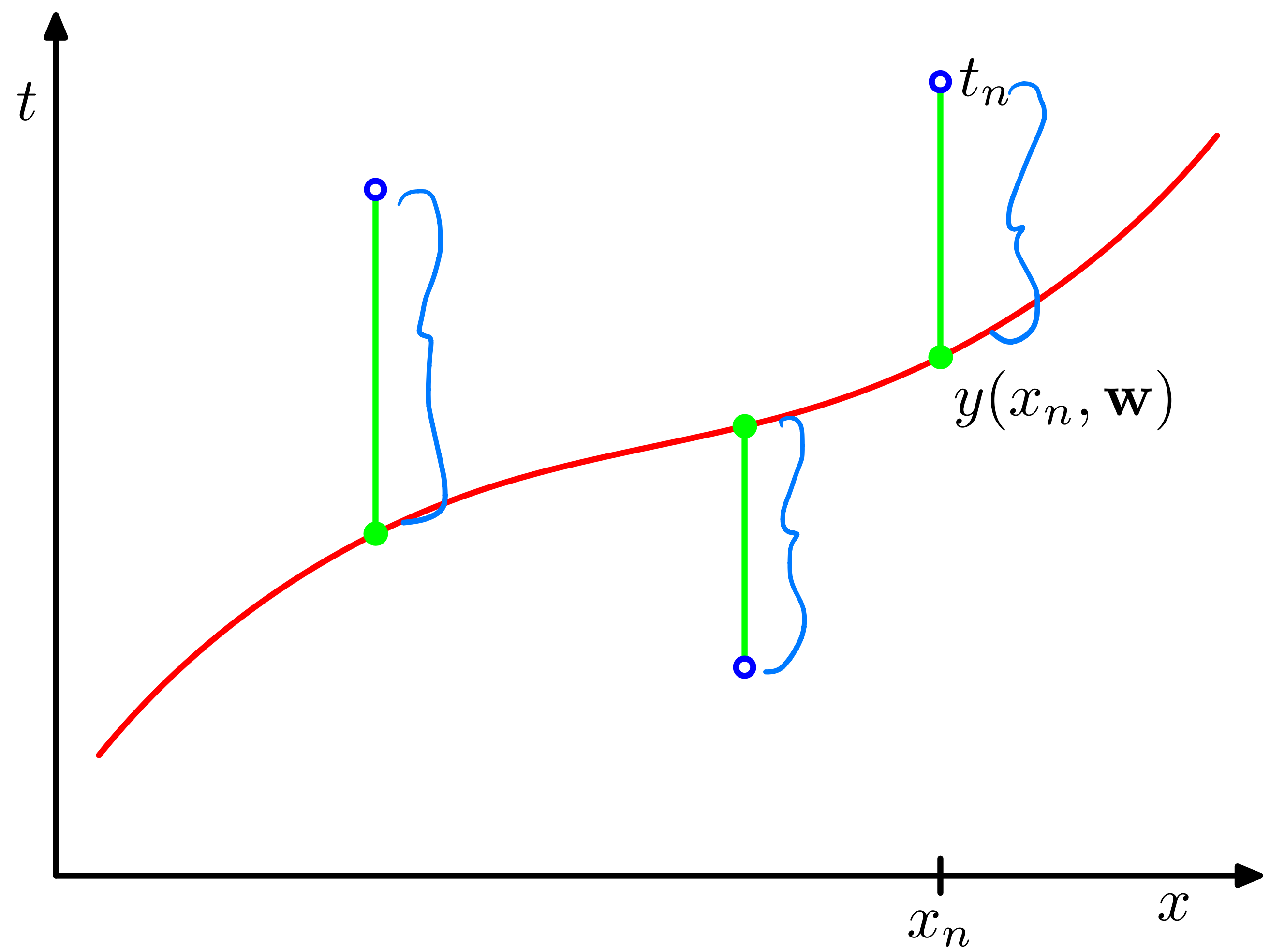
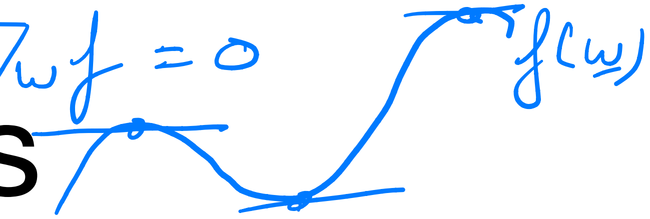


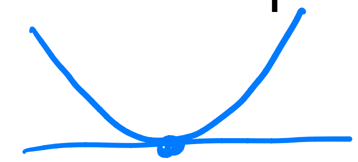
Figure: Errors are given by half the squares of green bars (Bishop 1.3)

Maximum Likelihood Estimates



- Maximize the log likelihood / Minimize the sum-of-squares error:

convex



$$\frac{\partial}{\partial \mathbf{w}} \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = -\beta \frac{\partial}{\partial \mathbf{w}} E_D(\mathbf{x}) = -\beta \frac{\partial}{\partial \mathbf{w}} \frac{1}{2} \sum_{i=1}^{N} \underbrace{\{t_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i)\}}^{2}$$

$$= -\frac{\beta}{2} \sum_{i=1}^{N} \frac{\partial}{\partial u} u^2 \frac{\partial u}{\partial w}$$

$$= + \frac{\cancel{\beta}}{\cancel{2}} \sum_{i=1}^{N} \cancel{2} \{t_i - \underline{w}^T \underline{\phi}(\underline{x}_i)\} \cdot \underline{\phi}(x_i)^T = 0$$

$\beta > 0$

$$\Longleftrightarrow \quad \underline{w}^T \sum_{i=1}^{N} \underline{\phi}(\underline{x}_i) \underline{\phi}(\underline{x}_i)^T = \sum_{i=1}^{N} t_i \underline{\phi}(\underline{x}_i)^T$$

$$\boxed{u = \{t_i - \underline{w}^T \underline{\phi}(\underline{x}_i)\}}$$

$$\frac{\partial u}{\partial \underline{w}} = -\frac{\partial}{\partial \underline{w}} (\underline{w}^T \underline{\phi}(\underline{x}_i))$$

$$= -\frac{\partial}{\partial \underline{w}} (\underline{\phi}(x_i)^T \underline{w})$$

$$= -\underline{\phi}(x_i)^T$$

verify

$\Longleftrightarrow$ (transpose)

$$\nabla_{\underline{w}} a := \frac{\partial a}{\partial \mathbf{x}} = \left(\frac{\partial a}{\partial x_1}, \frac{\partial a}{\partial x_2}, \ldots \right)$$

$$\left(\sum_{i=1}^{N} \underline{\phi}(\underline{x}_i) \underline{\phi}(\underline{x}_i)^T \right) \underline{w} = \sum_{i=1}^{N} t_i \underline{\phi}(\underline{x}_i)$$

Maximum Likelihood Estimates

design matrix

- Optimal $\boldsymbol{w}^*$ satisfies

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \dots & \phi_{M-1}(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \dots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

N x M matrix

vector of size M

verify

$$\sum_{i=1}^{N} \overbrace{\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T}^{M \times M \text{ matrix}} \mathbf{w} = \sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i) t_i$$

$$\Phi^T \Phi \underline{w} = \Phi^T \underline{t}$$

$$\underline{w} = \underbrace{(\Phi^T \Phi)^{-1} \Phi^T}_{} \underline{t}$$

Pseudo inverse

Φ^+ Moore-Penrose inverse of Φ

$(\Phi^+ \Phi = I)$

$$\mathbb{E}[t'|\mathbf{x}', \mathbf{w}_{\mathrm{ML}}] = \underline{w}_{ML}^T \phi(\underline{x}')$$